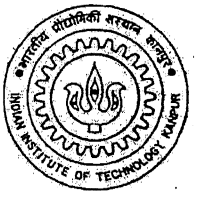
Spatio Temporal Concept Acquisition from Video with Commentary

A Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Master of Technology

by

V Shreeniwas



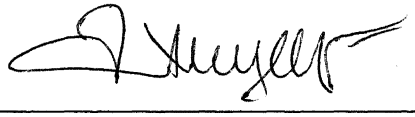
to the

Department of Computer Science & Engineering
Indian Institute of Technology, Kanpur

August, 2004

Certificate

This is to certify that the work contained in the thesis entitled " *Spatio Temporal Concept Acquisition from Video with Commentary*", by *V Shreeniwas*, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.



August, 2004

(Prof. Amitabha Mukerjee)
Department of Computer Science & Engineering,
Indian Institute of Technology,
Kanpur.

15 OCT 2004 CSE
पुरुषोत्तम नारायण हेगड़े पुस्तकालय
भारतीय प्रौद्योगिकी संस्थान कानपुर
अवाप्ति सं० A149184



Abstract

Computational models of language acquisition need a linguistic corpus synchronized with the same visual scenes. In this work, we use well-known psychological films along with commentaries by human subjects, to meet this need. We show how this can be used to identify event structures associated with verbs like "chase", "hit", or "move away". Some of the important aspects of this work are:

a. How to identify a feature set consistent with image schema such as path and animacy, which are likely to be well formed in children;

b. How to use these features to train a learning module.

Using Simple Recurrent Networks, we find that even from commentaries based on a single video, significant constraints are imposed on the possible interpretations of action words. We also show how the errors are reflective of early language acquisition, and how feedback results in further semantic convergence.

Acknowledgements

I am grateful to my thesis supervisor Prof Amitabha Mukerjee for his support and guidance through out the duration of this research work. I would like to thank Bridgette Martin and Barbara Tversky for providing us with the data and crucial inputs to our work, and to the organizers of the International Spatial Cognition Summer Institute (ISCSI) 2003 for providing the opportunity to interact with this group and discuss this data. My gratitude to Pradeep Vaghela for being a good partner in the lab and in my research work. I would also like to thank Dana Ballard, Prithwijit Guha, Pabitra Mitra, Achla Raina, G. Rajesh, V V Saradhi, KS Venkatesh, and the NLP and Vision research groups at IITK, for feedback and comments on the work discussed here.

I am also thankful for the staff in the Department and the rest of the Institute for being supportive during my entire stay and making it easy and enjoyable. I thanks my friends and well wishers for making my stay memorable. And last but not the least, I thank my parents for supporting me through every inch of my career, through all the thorns and flowers, through all successes and failures. I thank all of you.

Contents

1 Introduction 1

1.1 Language Acquisition 2

1.2 Time Series Classification 4

2 Capturing Verbs from Commentary 6

2.1 Synonym Detection 7

2.1.1 Considerations for Clustering 9

2.1.2 Final set of Verbs used for the study 9

3 Visual Feature Perception 11

3.1 Feature Selection 11

3.2 Features used: 12

3.2.1 Monadic Features 12

3.2.2 Dyadic Features 13

4 Learning Process 15

4.1 Elman Network 16

4.2 Environment 18

4.3 Results 18

4.4 One-circuit-for-one-Verb Framework 20

4.5 One-circuit-for-one-Verb Framework from Synthetic Data 23

4.6 One-circuit-for-all-Verbs Framework from Synthetic Data 25

5 Feedback Correction to reduce Focus Mismatches 30

6 Conclusion and Future work 32

A Sample Sequences from Synthetic Video 34

B Resolution of Lexical Units 36
B.1 Original List 36
B.2 Clustering on common Cognate 37
B.3 Final set of verbs used for the study 38

C Feature Extraction Methodology 39
C.1 Static-Features Extraction 39
C.2 Temporal-Features Extraction 41
C.3 Dynamic Tracking system to handle Occlusion, Contact and New Entries 42

Bibliography 43

List of Tables

2.1 Alternate Views of two subjects [*Chase* domain] 8
2.2 Final List of Verbs 10
3.1 Monadic Features 12
3.2 Dyadic Features 13
4.1 Percentage results using first framework 23
4.2 Interval results using One-Circuit-for-One-Verb Framework 23
4.3 Percentage results obtained through second Framework 24
4.4 Frame-by-frame results for One-Circuit-for-all-Verbs Framework Each Video were 2530 frames (81 seconds) each. 26
4.5 Interval classification results for One-Circuit-for-all-Verbs Framework 26
5.1 Improvement after feedback correction of *move away* verb 31
B.1 Original List of Predicates in *Hide and Seek* domain 36
B.2 Original List of Predicates in *Chase* Domain 37
B.3 Final List of Predicates 38

List of Figures

1.1 *Performance of learning system after six commentaries* Top row shows time intervals identified by humans as the evtn, and bottom row shows results from our system, VICES. The system performs well with simple verbs such as *moves*, while it does poorly with *Chase*. With successive refinements, involving additional data and error feedback, the performance is vastly improved. 2

2.1 Sample from *Chase* video for the interval 840-1110 frames. The corresponding description from the first subject in table 2.1: *The big square hit the little square again; the little circle moves to the door; the big square threatens the little circle.* 7

3.1 Illustration of value of Chamfer for some sample cases: (A) = 1; (B) = 0; (C) = $1/\sqrt{(2)}$; (D) = 1/2; (E) = -$1/\sqrt{(2)}$; (F) = 0; 14

4.1 Conceptual illustration of Elman Network 16

4.2 Samples of intervals where VICES performs a *Focus Mismatch.* As can be observed, in the chase case, the event between small square and circle closely resembles *chase.* Similarly, the other actions also closely resemble the corresponding semantics. 20

4.3 Sample intervals used to train VICES in first framework 22

4.4 shows Sample intervals where VICES correctly Classifies an event. . . 24

4.5 These figures compares the descriptions of the human subjects and that of VICES. The odd row bars indicate intervals where human talk of the event and the even row bars indicate the intervals where VICES detects the event. Black indicates true positives, dark Grey indicates false positives and light Grey indicates Focus Mismatches. False negatives are indicates by the absence of any bar below a black bar in the "Human" row. In each action, each set of human-VICES row is for one pair of objects among the 6 pairs possible. 27
4.6 compares the descriptions between humans and VICES for monadic verbs. 28
4.7 plots the performance of VICES with respect to human descriptions. The same pair of objects are chosen in these plots as are in fig 4.5 . . 29

5.1 A comparison of the output of VICES before and after feedback correction for *move away* verb. 31

A.1 Samples of sequences used to provide additional data to train VICES. 34
A.2 Samples of sequences used to provide additional data to train VICES. 35

B.1 Clustering of lexical units in *Hide and Seek* domain 37
B.2 Clustering of lexical units in *Chase* domain 38

C.1 Connected Component Analysis. A shows the initial set of pixels. B indicates the numbering after the first pass - horizontal assignment. C indicates the numbering after second pass - vertical assignment. D shows the final numbering after third pass - diagonal assignment. . . 40
C.2 Theta Computation for squares and rectangles 41
C.3 Track Box for Dynamic Tracking 43

Chapter 1

Introduction

"Bull in pure form is rare; there is usually some contamination by data."

-William Perry.

Consider a one year old child who is observing a scene along with adults who are talking about the scene. She is able to distinguish word boundaries and mark the repeated presence of certain words that are co-temporaneous with certain objects (nouns) or certain actions (verbs). From this, she makes inferences about these words and their meanings.

In this work, we consider an exemplar of this phenomenon with data that is extremely simple and direct, what may be called "an instance of laboratory purity." The "scene" is a set of 2D blobs moving around and the "commentary" are textual descriptions by humans who nonetheless exhibit comments invoking emotions such as "bash", "confuse" etc. The film on which this was based was created by Heider and Simmel (1944), which is a startling piece of socio psychological analysis which retains its research significance even today [20].

After listening to such commentaries from six adults, the child forms a partial understanding of the visually grounded meaning of verbs like "moves" or "chase" as in fig 1.1.

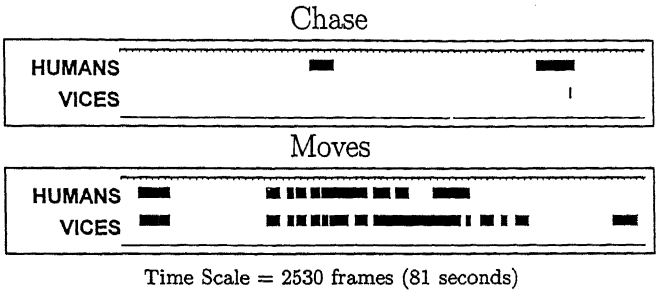


Figure 1.1: *Performance of learning system after six commentaries* Top row shows time intervals identified by humans as the evtn, and bottom row shows results from our system, VICES. The system performs well with simple verbs such as *moves*, while it does poorly with *Chase*. With successive refinements, involving additional data and error feedback, the performance is vastly improved.

1.1 Language Acquisition

Learning the meanings of words is a difficult inductive problem; when a new word is encountered in a sensory context, it is not clear which aspect of the scene it refers to [31]. Even after the same object is encountered many times, it is not clear whether the inductive generalization is complete [17]. Despite this, children are able to constrain the possible interpretations of a word over repeated exposure.

Computational models of this process of language acquisition acquired considerable importance, not only because of the potential of the technology itself, but also because of its role in synthetic psychology. However, attempts in learning grounded semantics have been limited by the availability of a corpus of language data corresponding to the identical sensory stimulus.

In this work, we take a video based on the well known video of moving geometric objects [13], on which a large number of human subjects have been asked for their commentary. Both the language commentary (in text format), as also the visual scene is repeatedly shown to the system, which eventually forms a preliminary model of the words. In particular, we focus on event structures associated with verbs, which have traditionally been under-investigated compared to nouns.

Thus, this process is analogous to repeating a scene to a child, along with commentary from a distinct adult each time. Since distinct adults speak each time, it is quite possible that the verbs used might be different, but related. Would the child be able to cluster such verbs into distinct categories based on temporal overlapping? What are the set of visual features that it shall require to recognize such events? What kind of processing abilities would it need? This thesis discusses a framework which tackles all these challenges and evolves a computational model for pre-linguistic event structure acquisition.

Research on recognition of temporal patterns (actions, events etc) from visual stimulus have been based so far only on visual inputs [44, 14, 23, 27, 28, 18]. Though such methods can be effectively used for indexing and information retrieval [48] , these approaches cannot be used for language acquisition.

Language acquisition using language input using recurrent circuits have been investigated [47, 16, 1, 5] but absence of visual grounding renders them ineffective for the task at hand.

Efforts to language acquisition from visual input alone or from multi modal input have relied on hand crafted knowledge of the semantic or logical structure of the language units [8, 41, 29, 42, 23]. Though [43] talks about physics notions for grounding language, it is not sure whether Counter-factual Simulation, the perceptual process used to recover the primitives, can be performed by pre-linguistic organisms. In effect, these approaches to provide systems with symbol-grounding ability heavily rely on hand crafted perceptual representation provided by AI designers.

Recent efforts by Roy have demonstrated language acquisition from multi modal input without an a priori knowledge base. However, they have concentrated primarily on Noun structure acquisition[34, 35, 2]. Steels demonstrates a method for building a language through shared learning [45]. The resultant language through this method may be novel every time, and is directed more towards language invention than language acquisition. None of these methods, therefore, give a satisfactory answer to automatic symbol grounding for verb structures from video.

It is established that children have certain pre-linguistic perceptual processing abilities that include the conceptual primitives of Path, Up-Down, Containment,

Force, Part-Whole and Link [19]. Based on these, Perceptual Analysis, or the process of attentively analyzing to abstract new kind of information, can be performed by children [19]. It has already been shown that the above abilities of a child can be feasibly built into artificial agents [26].

In this work, we use two psychological research films, one based on the classic Heider and Simmel's (1944) *Chase* video and another based on the game of *Hide and Seek* as the visual grounded corpora. We use commentaries, from independent human subjects, describing actions at a fine level, as the linguistic input. Based on semantic clustering of these commentaries, we extract verbs. These form the second input to the system. Based on a set of visual features and processing abilities, both argued to be existent in children, we form a computational model for acquiring event structures. The structures learned are in the form of weights of Simple Recurrent Networks (SRNs) and therefore can be considered as an analog form of the event structure. We call our system VICES - Video and Commentary for Event Structure acquisition.

Going from a small network that can acquire a single event structure from limited data, we successively refine the framework to achieve a single structure that can distinguish between 6 different verbs.

1.2 Time Series Classification

There are many tools such as Hybrid ANN-HMM, PERUSE, Neural Networks and SVMs that have been established for temporal processing. They can learn temporal patterns and are hence capable of predicting unseen members of temporal sequence, but cannot perform time series classification[9, 25, 24, 37, 36]. Complex Bayesian methods (including HMMs) or probabilistic classifiers have been proved to do a good job of recognizing complex actions in video [21, 3, 40, 27, 28], however, there is little evidence of such advanced inference abilities in children. Only HMM and Recurrent Neural Networks have been shown to have abilities to perform *Time Series Classification.* HMMs make very large assumptions about data, require much larger training data, and the number of parameters that need to be set in an HMM are

huge. Moreover, HMMs only learn from positive exemplars and the number of states and transitions are often set only by an educated guess [15]. On the other hand, Recurrent Neural Circuits do not suffer from any of these drawbacks, but are limited by the amount of past context they can consider. Certain modifications of SRNs, such as LSTMs, are capable of considering a longer context [39].

There is evidence that the Cerebral Cortex can be modeled using a SRNs [32, 12, 10]. Recurrent Neural Networks have also been used in the domain of Language Acquisition to infer well-formedness of sentences, represent lexical categories and token distinction, predicting lexical categories, and constructivist language acquisition [16, 1, 5, 47, 7, 6].

Based on the discussion on visual features in [46], we have selected a set of features, that gives maximum distinction between the verbs involved in the study, and the informativeness thus provided by the feature. These features are consistent with the discussion provided in [33].

The Bayesian framework, provided in this work, in combination with these visual features aims to provide not only a mathematical model to solve abstract perceptual problems under uncertainty, but also a plausible model of how biological systems might work.

Chapter 2

Capturing Verbs from Commentary

This work uses a corpus of commentary on two well-known videos collected in a psychological experiment conducted by Martin and Tversky [20]. Visual stimulus and linguistic descriptions are provided simultaneously as the multi modal inputs to the event structure acquisition process. The experiment involved showing two dimensional movies to human subjects; their commentary was recorded and transcribed. The experiment used two animations portraying motion paths of geometric figures: one was an adaptation based on the classic Heider and Simmelś (1944) *chase* and the other based on the game of *hide-and-seek*. Participants were asked to segment the videos based on boundaries of perceived actions, while simultaneously providing commentary on the movies. Details of the experiment, the videos and the results can be found in [20, 49].

As part of their experiment, Martin and Tversky segmented the commentary into sentences, the majority of which contained one verb, resolved anaphora, and assigned the statement to one of several segments. For the purpose of this study, only fine level descriptions in both the domains were selected, with the video running forward in time.

The segmentations of the video were converted into frame references which were linked to the statement recorded in the corpus. There were a total of 16 subjects

presenting a total of 309 statements. Sample sets of transcripts, from two subjects are provided in table 2.1. A sample of the video for the interval 840 - 1110 is provided in fig 2.1.

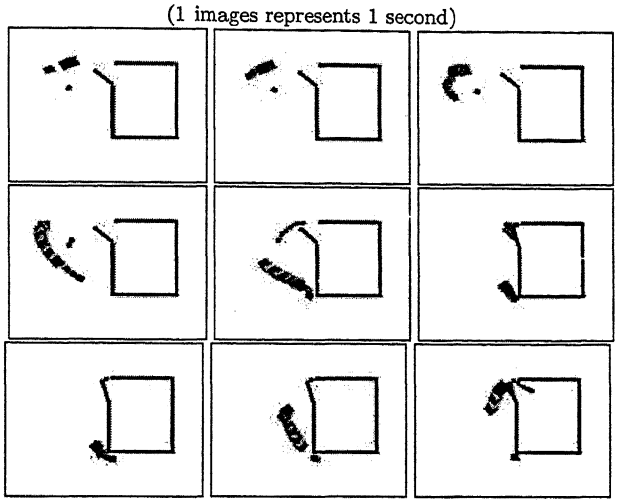


Figure 2.1: Sample from *Chase* video for the interval 840-1110 frames. The corresponding description from the first subject in table 2.1: *The big square hit the little square again; the little circle moves to the door; the big square threatens the little circle.*

2.1 Synonym Detection

Classically, language acquisition in the child proceeds with an initial competence in vocabulary (Stage I) followed by acquisition of grammar (Stage II) [30]. However, there is evidence that although language generation uses a weaker syntax, syntactic cues are used in sentence comprehension at a much earlier stage [22]. Thus, there is evidence that some degree of syntactical competence is plausible in language learners in Stage I. For the purposes of this experiment where we are constrained to 309 utterances, this is a practical necessity, for it would not have been possible to detect all morphological variants of each verb. Thus, it is assumed that morphological variants (chases, chased, were chasing) are accessible, as are active-passive variants (chases vs be chased by).

Start Frame	End Frame	Subject One	Subject Two
10	57	the door closed	the square closes the other, the smaller square
172	177	The square and the circle move into the screen	two more objects came in, a circle and a square
387	398	the big square moved out of the corner	the square is moving around
487	540	the big square opened the door	he's trying to push his way out of the square
617	635	the little square hit the big square	they're hitting each other
805	848	the big square hit the little square	and they keep hitting each other
852	1100	the big square hit the little square again; the little circle moves to the door; the big square threatens the little circle	now the circle is blocking the entrance for the big square; now the circle is inside the square
1145	1202	the big square goes inside the box; (and) the door closes	another square went inside the big square
1270	1630	the big square approaches the little circle in the; the little square opens the door	the little square's trying to get inside the big square
1550	1658	the square scares the little circle in the corner	and the objects inside are moving closer together
1752	1796	the little circle goes out of the room	the circle just went out of the square
1780	1853	the door closes	the little circle, the little square closes box
1967	1996	the big square hits the door down	the big square just got out of the box
2197	2207	the little square and the circle go off the screen	they left the picture
2207	2292	the big square circled around	now the big square is there with the square box
2471	2501	big square knocks the door down	(the big square) tried to force his way in

Some clauses have for one subject have been combined to synchronize with the clauses of the other subject

Table 2.1: Alternate Views of two subjects [*Chase* domain]

The verbs collected at the end of the experiment outlined needs to be clustered before proceeding into the structure acquisition phase. The need for clustering arises because of the involvement of multiple subjects in the commentary collection process and the use of multiple lexical units by each subject. Multiple subjects describe the same events using different lexical units. The acquisition process would not be able to distinctly learn so many structures synonymous with each other with the limited data available to us.

Therefore, the lexical units are analyzed, some infrequent ones are eliminated, and others are replaced with their synonyms.

2.1.1 Considerations for Clustering

The clustering is done on the basis of context specific semantic relatedness [4]. Though there are many methods used to measure and compare Semantic Relatedness, we use the simple measure of user evidence. Human judgments yield the most generic assessment of the goodness of a measure [4]. The criteria used for the resolution are listed below:

1. *Overall Frequency:* Lexical units with frequency of just one were removed as being too infrequent for our purpose.

2. *User Evidence of Synonymy:* Any pair of lexical units that were used consistently in the same interval, to mark the same action, for the same set of agents, by two different subjects were considered as synonymous. In case of multiple such pairs involving the same event, the one with highest frequency was considered. Among each pair, the less frequent lexical unit was eliminated and the more frequent one was retained. We encountered two types of such cases:

 (a) *Overall Synonymy:* In this case, the less frequent lexical unit was used by all subjects, but the overall frequency was still very less.

 (b) *Frequency across users:* In this case, the less frequent lexical unit was used by only a small minority of the subjects and the remaining majority of the subjects used different lexical units.

Details of this process are given in Appendix B.

2.1.2 Final set of Verbs used for the study

After the process of similarity clustering, the final list of verbs, with their semantic connotations and argument structure, are given in table 2.2.

Verb	Arguments	Semantics
chases	X, Y	X is chasing Y (X behind, Y ahead)
come together	X, Y	X and Y come together to each other
move away	X, Y	X and Y move away from each other
hit	X, Y	X hits Y
moves	X	X moves
spins	X	X is spinning

Table 2.2: Final List of Verbs

Chapter 3

Visual Feature Perception

In this section, the considerations behind choosing the set of features used in this work, are discussed. The features and the formulae used to compute them are briefly explained.

3.1 Feature Selection

There is evidence that children and lower organisms can form perceptual prototypes by abstracting the central tendencies of perceptual pattern [19]. Such conceptualization about events and temporal structure of events comes from observation rather than from physical interaction with objects [19].

One of the foundations of the conceptualizing capacity is the image-schema, in which spatial structure is mapped into conceptual structure. Image-schema are notions such as Path, Up-Down, Containment, Force, Part-Whole and Link, notions that are thought to be derived from perceptual structure [19] .

Both biological and artificial use perceptual systems that can extract certain features from a sensory array and use these features for categorization tasks, recognition and motor control. Several characterizations of the notion of "feature" have been proposed so far in specific fields of vision science. "Goodness" of featural information may be defined either through effectiveness in verb detection or may be defined as a function of reliability, and informativeness [46, 33].

Keeping these considerations in mind, we have selected a set of features, that gives maximum distinction between the verbs involved in the study, and the informativeness thus provided by the feature. All features selected are either kinematic features or are derived from them through temporal derivations or from applying a transform on the other features.

Features selected for our purposes are related to spatial aspects in image-schema such as position, relative pose, relative velocity etc.

Since the purpose of the study is to concentrate on event structure acquisition, noun structures with relation to the video used, such as "big square", "small square" etc are assumed to exist in the system.

3.2 Features used:

The features that are selected for the purpose of classification of events fall into two categories:

3.2.1 Monadic Features

Monadic features are those that pertain to only one object. These may either be used to detect Monadic verbs or for the computation of Dyadic features. The same are listed in table 3.1.

Name	Description	Formula
v_x	Velocity of the object in x direction.	$\frac{dx}{dt}$
v_y	Velocity of the object in y direction.	$\frac{dy}{dt}$
v	The velocity of the object.	$\frac{ds}{dt}$
a_x	Acceleration of the object in x direction.	$\frac{d^2x}{dt^2}$
a_y	Acceleration of the object in y direction.	$\frac{d^2y}{dt^2}$
a	Acceleration of the object.	$\frac{d^2s}{dt^2}$
$d\theta$	Angular displacement of the object.	$\theta_n - \theta_{n-1}$
ω	Angular velocity of the object.	$\frac{d\theta}{dt}$
α	Angular acceleration of the object.	$\frac{d^2\theta}{dt^2}$

All differentials are computed over a one frame interval.
* To counter aliasing, a sliding window of 11 frames is used to compute this. This gives us a smooth transition in this variable.

Table 3.1: Monadic Features

3.2.2 Dyadic Features

Dyadic features are those that pertain to two objects taken together. These features are informative about the relative dynamics between the two objects. They are used for the detection of dyadic verbs. The same are listed in table 3.2.

Name	Description	Formula
proximity	It is inverse of the boundary-to-boundary distance between two objects.	$\frac{1}{r-(\sqrt{m_1}+\sqrt{m_2})}$
relVel	Relative Velocity between the two objects	$\frac{dr}{dt}$
relAcc	Relative Acceleration between the two objects	$\frac{d^2r}{dt^2}$
Parallelism Measure	Measure of Parallelism between the direction of motion of the two objects	$\cos(\hat{v_a} - \hat{v_b})$
Leader A	Measure of leadership (in motion) of A w.r.t B	$\cos(\hat{v_a} - \theta_{ba})$
Leader B	Measure of leadership (in motion) of B w.r.t A	$\cos(\hat{v_b} - \theta_{ab})$
Chamfer	Measure of chamfering between the two objects	$\frac{\sin(\hat{v_a}-\theta_{ba})+\sin(\hat{v_b}-\theta_{ab})}{2}$

where:
r : inter CG distance.
θ_{ab}: angle of vector between $CG_a -> CG_b$

Table 3.2: Dyadic Features

An explanation on the dyadic features (table 3.2) is mandated at this point. High *proximity* indicates good spatial clustering between these two objects at this instance of time. It provides the learning algorithm a method to select pair of objects for detection of dyadic verbs among all pairs of objects.

Relative Velocity and *Relative Acceleration* provide measures of the direction of relative motion of these objects. It clearly distinguishes between the scenarios when the objects are moving towards each other or are going farther from each other.

Parallelism measure is an indication whether the two objects are moving in the same direction or not. It gives a high value when they are moving together in the same direction and a lower value when they are heading in different directions. It is also a measure of temporal clustering of the direction of the two objects.

Leader measures indicate leader or follower in chasing/following scenarios. Its measure is close to 1, when the respective object is ahead of the other object, 0 when the two are moving parallel and -1 if the object is behind the other object in relative motion.

Chamfer literally means any two surfaces meeting at any angle other than right angle. Chamfer measure indicates if the two objects are moving together or are

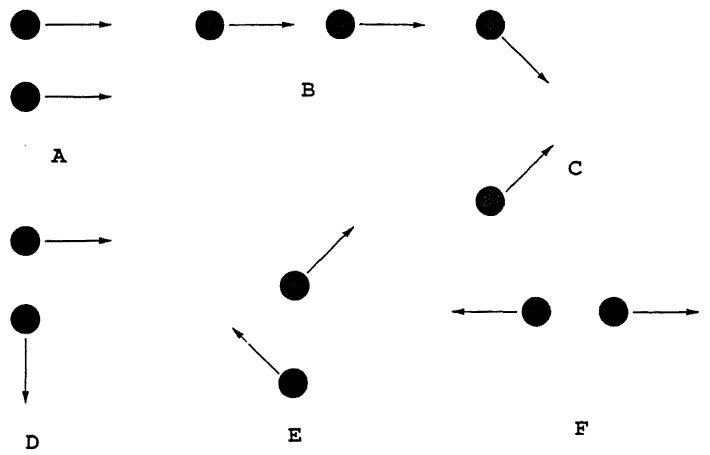


Figure 3.1: Illustration of value of Chamfer for some sample cases:
(A) = 1; (B) = 0; (C) = $1/\sqrt{(2)}$; (D) = 1/2; (E) = -$1/\sqrt{(2)}$; (F) = 0;

moving ahead of one another. An illustration of Chamfer is given in fig 3.1.

The feature Extraction Methodology is detailed in Appendix C.

Chapter 4

Learning Process

Once we have extracted our features over the entire video and we have the corresponding classification for the video, the learning problem that we face is one of Time Series Classification. Let us assume that there exists a mapping

$$y = f(x|x = d_1, d_2, d_3, ..., d_n)$$

where d_i is feature data at time instance i, y would then be the class of verb which this particular interval in the video visualizes. Given a sequence of Time series data and the corresponding classifications, we need a learning system that can learn this mapping. Given this mapping and a new time series, the problem of detection of verb is one of Time Series Classification.

Simple Recurrent Network (SRN) is an extension of the popular Neural Network architecture involving memory and hence is well suited for processing of temporal data. The appropriate response at a particular point in time depends not only on the current input, but potentially all previous inputs. A general taxonomy of neural net architectures for processing time varying patterns is given in [24]. SRNs give us the benefits of being able to process temporal data, with simple but powerful mathematical properties, and being able to learn from less data.

In SRNs, a "context" layer is added to the structure, which retains information between observations. At each time-step, new inputs are fed into the SRN. The previous contents of the hidden layer are passed into the context layer. These then feed back into the hidden layer in the next time step. In an algorithm similar

to the backpropagation algorithm, called back propagation through time (BPTT), the weights of the hidden layers and context layers are set. To do classification, postprocessing of the outputs from the SRN is performed; so, for example, when a threshold on the output from one of the nodes is observed, we register that a particular class has been observed.

4.1 Elman Network

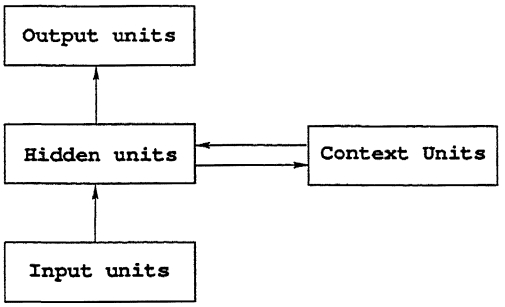


Figure 4.1: Conceptual illustration of Elman Network

The Elman Network, a specific type of Recurrent Neural Network Architecture, now referred to as Simple Recurrent Network (SRN), is commonly a two-layer network with feedback from the first-layer output to the first layer input. This recurrent connection allows the Elman Network to both detect and generate time-varying patterns [5]. A two-layer Elman Network is shown in fig 4.1. The Elman Network has Tan Sigmoidal *(tansig)* neurons in its hidden (recurrent) layer, and Linear *(purelin)* neurons in its output layer. Tansig neurons can output values between -1 and 1, while purelin can take on any value. If purelin neurons are used in the output layer, the network outputs can take on any value. This combination is special in that two-layer networks with these transfer functions can approximate any function (with a finite number of discontinuities) with arbitrary accuracy. The only requirement is that the hidden layer must have enough neurons. More hidden neurons are needed as the function being fit increases in complexity.

The Elman Network differs from conventional two-layer networks in that the first layer has a recurrent connection. The delay in this connection stores values from the previous time step, which can be used in the current time step. Thus, even if two Elman Networks, with the same weights and biases, are given identical inputs at a given time step, their outputs can be different due to different feedback states. Because the network can store information for future reference, it is able to learn temporal patterns as well as spatial patterns.

Because Elman Networks are an extension of the two-layer sigmoid/linear architecture, they inherit the ability to fit any input/output function with a finite number of discontinuities. They are also able to fit temporal patterns, but may need many neurons in the recurrent layer to fit a complex function.

There is evidence in literature that SRN architectures can be used to perform Bayesian inference for arbitrary hidden Markov models. This suggests that such neural structures can be used to model cerebral cortex [32]. Neural mechanisms have also been developed to recognize biological morion compatible with known neuro-physiological facts [10, 12, 11].

Since the data available through the psychological experimental videos is very less (It consisted of only one video of 2530 frames, in each domain, with intermittent intervals of truth of these verbs), we present three different frameworks:

1. **One-circuit-for-each-Verb - using language input** In this framework, we build one neural circuit for each of the verbs and train it with the available video and commentary. With this, we establish that such a framework can effectively perform perceptual analysis and evolve an analog form of the schema used to represent the particular verb. Since the data was limited, we could not train a single Neural Network to recognize all the verbs.

 In this case, the feature vector used for the dyadic verbs consist of the monadic features of both the objects and the dyadic features for each pair. In case of monadic verbs, only the monadic features of the concerned object is used.

2. **One-circuit-for-each-Verb - using synthetic data** Since the data available is so less, we generate synthetic data in the form of a video animations

akin to the movies used earlier. We use this synthetic data to train the SRNs in the same way as in the previous framework. It contains multiple exemplars of each verb of the following verbs - *chase, follow, move away, come closer, walk* and *run.* And the script for this animation serves as a manual verb annotation for the purpose of learning, and therefore, a commentary is not required. Sample sequence of each verb from the video is provided in Appendix A. We demonstrate the additional data to prove that the performance of the model can be improved by using additional data.

Here, in the case of dyadic verbs, the feature vector consists of only dyadic features of the pair.

3. **One-circuit-for-all-Verbs - using synthetic data** In this framework, we use the synthetic video to generate one neural circuit capable to distinguishing among all the verbs. We then again test it against the same classes to demonstrate that such a network is feasible to construct and train, provided there are ample exemplar to learn from.

 Here, in the case of dyadic verbs, the feature vector consists of only dyadic features of the pair.

4.2 Environment

The program to extract the Features is programmed in Java. It generates the feature vectors for each frame in the video. The SRN training and testing is performed in Matlab. The environment included a Intel Pentium IV 900 MHz processor with 192 MB RAM and 20GB HDD running Microsoft Windows 2000. Further data requirements were satisfied by using local FTP space.

4.3 Results

We use two related, but distinct measures of performance as given below. For each verb:

1. *The frame to frame accuracy of the results:* This is the percentage of the correctly classified frames as against the total number of frames.

2. *The interval detection accuracy:* This is a percentage of the correctly classified intervals as against the total number of intervals.

The first measure is indicative of the performance of the SRN on a frame by frame basis. It is a good measure of whether the SRN delimits the truth of the verb as soon as it begins and ends with respect to human transcription. The length of the intervals can play a crucial role in this measure. While the SRN might be able to detect an interval, it might be able to classify it only for a small part of the interval - in which case, the accuracy will be negatively affected.

Also, since the SRN is simulated for the entire video, it is quite often possible that it may take time to switch from one verb to another. Hence, while VICES may rightly classify an interval as a whole, but it may not be able to align the detected verb alongside the exact interval. If that happens, then the second measure will be higher than the first measure. On the other hand, while the system may not be able to detect the entire interval, it may be able to correctly classify the verb at intermittent, shorter intervals. In this case, the first measure will be better than the second one. Therefore, both these measures are indicative of the performance of the system. Ideally, a system that can perform better on both measures would be highly appreciated. . .

While testing the system, some of the false positives are actually events very similar to the event in consideration. There are intervals where an event may be occurring, however, the human subjects do not notice or describe it because their attention is focused on something else. We name them as **Focus Mismatches**, since the mismatch from the expected output is because the system does not have any notion of focus at this point of time. An example situation is when A is chasing B & C, there might be a chase-like action occurring between B and C, which is generally ignored by the subject. These focus mismatch situations are separately identified in the results.

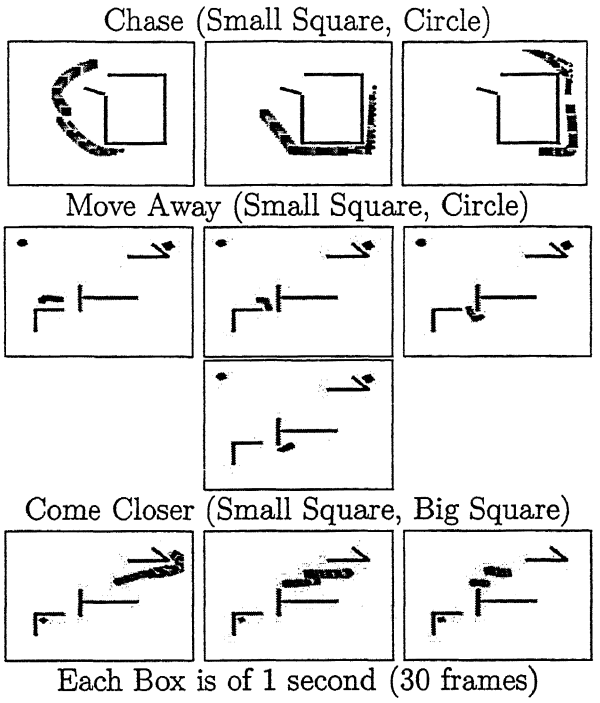


Figure 4.2: Samples of intervals where VICES performs a *Focus Mismatch.* As can be observed, in the chase case, the event between small square and circle closely resembles *chase.* Similarly, the other actions also closely resemble the corresponding semantics.

4.4 One-circuit-for-one-Verb Framework

There are two inputs to the learning system - the feature vectors at each instance of time and the corresponding action verb.

For each action verb:

- We provide the learning system with the intervals where it is true and some other intervals where the verb is false.
- Each frame where the verb is true is marked as 1 and 0 otherwise. This forms the *desired output* vector. The *input* is an array of feature vectors, one for each frame.
- Based on this, one SRN is trained with this *input* and the *desired output.*

These Neural Networks are trained for 600 epochs of learning, or convergence, whichever is the sooner. None of the networks reach convergence, but achieve Average Error of about 1%.

- These SRNs had one input neuron for each feature, same number of hidden neurons as in the input layer and one output neuron.
- To test these SRNs, it was simulated over features of video sequence it had not seen during training. We test the SRNs on all possible cases out of these two videos that was not shown during the training process. We also make a conscious attempt to provide as little training data as possible, because it would then give more data to test these SRNs.

In table 4.3, illustrations of the examples used as training are provided.
Let the total time of visual sequence for each verb be t time units.

- E : Intervals when subjects describe an event as occurring.
- $\overline{E} : t \setminus E$
- E' : Intervals when VICES describes an event as occurring.
- $\overline{E'} : t \setminus E'$
- FM: Intervals classified as *Focus Mismatches.*

$$\text{True Positives} = \frac{E \cap E'}{E}$$

$$\text{False Positives} = \frac{E' \setminus E}{\overline{E}}$$

$$\text{False Negatives} = \frac{E \setminus E'}{\overline{E}}$$

$$\text{Focus Mismatches} = \frac{FM}{\overline{E}}$$

$$\text{Accuracy}^1 = \frac{(E \cap E') \cup (\overline{E} \cap \overline{E'})}{t}$$

[1]During Accuracy computation, Focus Mismatches are included in False Positives

भारतीय प्रौद्योगिकी संस्थान कानपुर
अवाप्ति सं० 149184

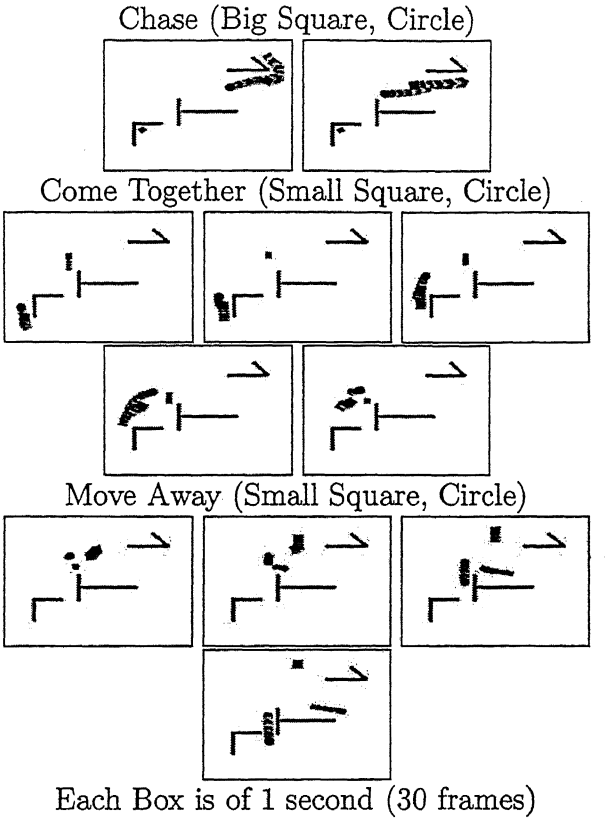


Figure 4.3: Sample intervals used to train VICES in first framework

Tables 4.1, 4.2 give the results using the first framework. In table 4.4, illustrations of those intervals which are correctly classified by the learning system are provided.

To assess the performance of the system, we plot the system's classification in comparison with the human descriptions over the time line of the video. Fig 4.5 shows plots showing the comparison. The rows of the plots are the various combinations of objects i.e small square - circle, big square - circle etc in the two videos. The times lines are provided for the monadic verbs in fig 4.6. In all time lines, dark Grey color indicates false positive classification, while light Grey color indicates focus mismatches.

The accuracy is lower in the case of verbs with more complex semantics (*hit, chase*) as against verbs with simpler semantics (*moves, spins*). However, the high

Verb	True Positives	False Positives	False Negatives	Focus Mismatches	Accuracy
hit	46.02%	3.06%	53.98%	2.4%	92.37%
chase	24.44%	0%	75.24%	0.72%	93.71%
come Closer	25.87%	14.61%	73.26%	16.77%	63.66%
move Away	46.34%	7.21%	52.33%	15.95%	73.37 %
spins	82.54%	0%	16.51%	24.7%	97.03%
moves	68.24%	0.12%	31.76%	1.97%	77.33%

Table 4.1: Percentage results using first framework

Verb	True Positives	False Positives	False Negatives	Focus Mismatches
hit	3	3	1	1
chase	6	0	3	4
come Closer	6	20	7	24
move Away	8	3	0	14
spins	22	0	1	9
moves	5	1	2	7

Table 4.2: Interval results using One-Circuit-for-One-Verb Framework

accuracy in the case of *spins* and *hit* can be attributed to very few occurrences of the event in the given data. Therefore, the SRNs tend to over specialize. In case of *chase* and *come closer*, false negatives percentage is higher. In case of *chase*, VICES detects some intervals, but for a shorter duration and therefore the false negatives are higher.

4.5 One-circuit-for-one-Verb Framework from Synthetic Data

The perceptual analytical abilities in child are suggested to be achieved through the process of observation and not through actual participation in the corresponding actions [19]. Based on this, it can be assumed that to learn the distinctions between semantically similar verbs such as *chase* and *follow* (distinguished on the basis of speed) or between *chase* and *run* (distinguished on the basis of relative positions during motion), the verb acquisition system would need more exemplar.

In this framework, we generate a video animations akin to the movies used earlier.

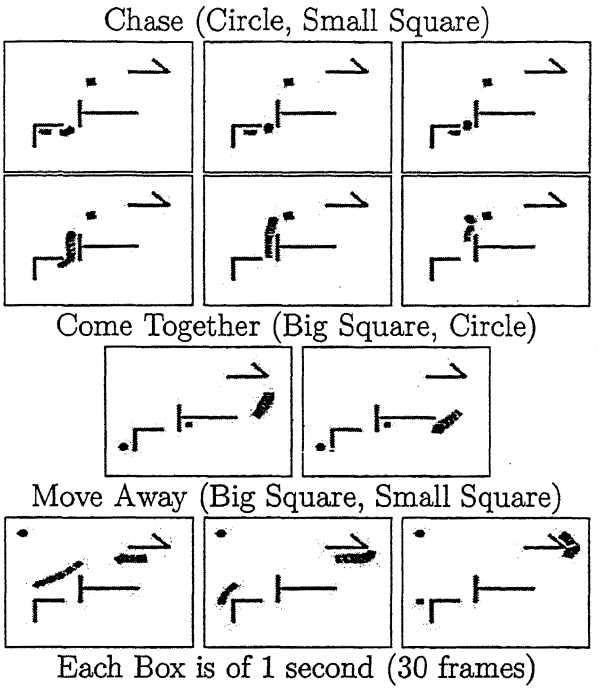


Figure 4.4: shows Sample intervals where VICES correctly Classifies an event.

It contains multiple exemplars of each verb. And the script for this animation serves as a manual verb annotation for the purpose of learning, and therefore, a commentary is not required. Sample sequence of each verb from the video is provided in Appendix A.

We then, develop one neural circuit for each of the verbs and train it using this video. We then test it against the same cases used in section 4.4. Table 4.3 shows the results obtained using this framework. The percentages are computed as before.

Verb	True Positives	False Positives	False Negatives	Focus Mismatches	Accuracy
chase	52.59%	0.21%	47.09%	1.24%	95.14%
come Closer	53.61%	11.29%	45.52%	20.41%	66.66%
move Away	65.30%	12.07%	33.37%	17.15%	70.29%

Table 4.3: Percentage results obtained through second Framework

Table 4.7 provide comparison on the time line, as in 4.4.

We establish that having more data certainly improves the classification accuracy. However, since the verb annotation for this video was done a human operator, the effect of focus mismatches would not be observable. Focus Mismatches increases due to this reason. Recall that Focus Mismatches are due to the perceptual process of an observer, where attention is maximized on only a subset of all the simultaneously occurring events and therefore, the observer may fail to describe or take note of it.

The accuracy of VICES through this framework improves after using synthetic data. In case of *chase* all the measures improves. In case of *come closer* and *move away*, though focus mismatches slightly increase, the accuracy vis a vis true positives improve vastly.

4.6 One-circuit-for-all-Verbs Framework from Synthetic Data

In the third framework, we use the video of case 2 to generate one neural circuit capable to distinguishing among all the verbs. We then again test it against the same classes to establish that such a network is feasible to construct and train, provided there are ample exemplar to learn from.

Based on these training samples, a single SRN is trained for all the dyadic verbs. The Neural Network has one input neuron for each feature, one output neuron for each verb and the same number of hidden layer neurons as there are in input layer. The Neural network also has one output neuron to indicate nothing. This is done following the concept of *winner takes all* so that the effective output of the network can be computed as the neuron which has the maximum output. Rest of the training process was the same.

Here, the percentages are computed differently. Since either one of the output neurons fire at any given frame, E' would be approximately equal to t. Therefore, we just make a mention of E' in terms of frames. Every verb detection by the circuit is annotated manually as being correct or wrong. Based on these, the total accuracy is computed as a ratio of correct verb detection to the total time scale. For interval

accuracy, however, it is computed as a ratio of correctly classified intervals to the total intervals for each pair of objects.

Tables 4.4 and 4.5 lists the performance of the system using second framework. The reasonably high accuracies in all the cases demonstrates that it is possible to construct a single neural circuit to distinguish and identify many different events, provided the exemplar is large enough for the framework to generalize from.

Video	Objects involved	Frames Classified	Correct Classification	percentage
Hide & Seek	Small Square, Circle	2492	2380	94.07
	Small Square, Big Square	2510	2440	96.44
	Big Square, Circle	2530	2028	80.16
Chase	Small Square, Circle	2334	2285	90.30
	Small Square, Big Square	2329	2156	85.22
	Big Square, Circle	2384	2207	87.23

Note: The percentage accuracy is computed over total number of frames and not on number of classified frames.

Table 4.4: Frame-by-frame results for One-Circuit-for-all-Verbs Framework Each Video were 2530 frames (81 seconds) each.

Video	Objects involved	Intervals	Correct Classification	percentage
Hide & Seek	Small Square, Circle	71	62	87.32
	Small Square, Big Square	44	38	86.36
	Big Square, Circle	70	62	88.57
Chase	Small Square, Circle	87	80	91.95
	Small Square, Big Square	83	78	93.98
	Big Square, Circle	76	72	94.75

Table 4.5: Interval classification results for One-Circuit-for-all-Verbs Framework

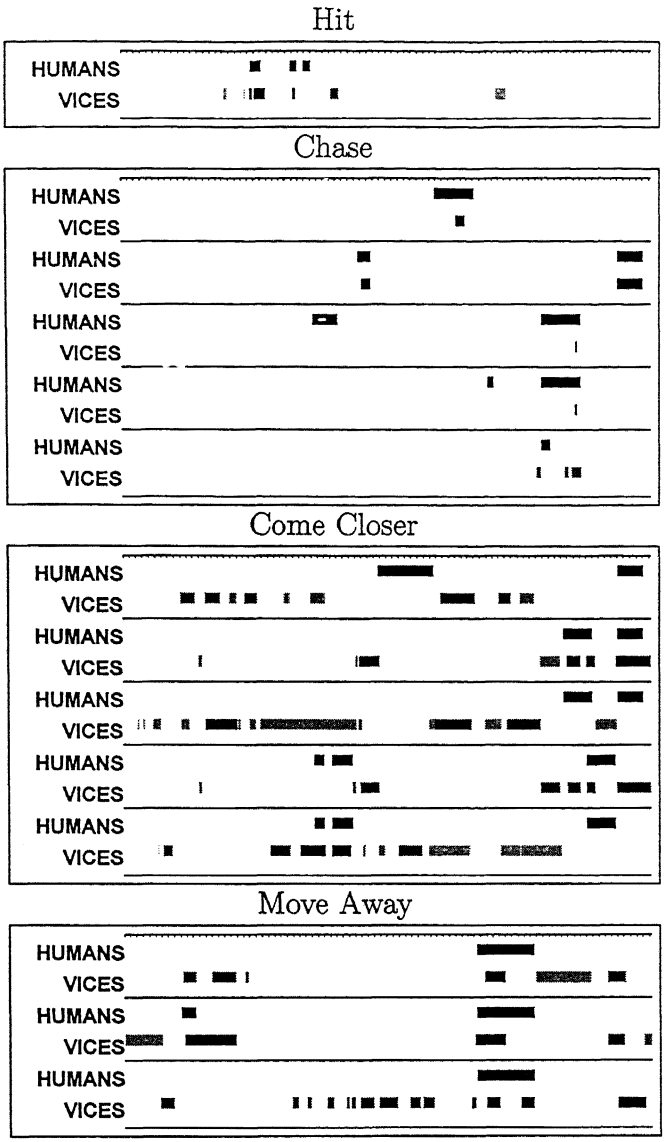


Figure 4.5: These figures compares the descriptions of the human subjects and that of VICES. The odd row bars indicate intervals where human talk of the event and the even row bars indicate the intervals where VICES detects the event. Black indicates true positives, dark Grey indicates false positives and light Grey indicates Focus Mismatches. False negatives are indicates by the absence of any bar below a black bar in the "Human" row. In each action, each set of human-VICES row is for one pair of objects among the 6 pairs possible.

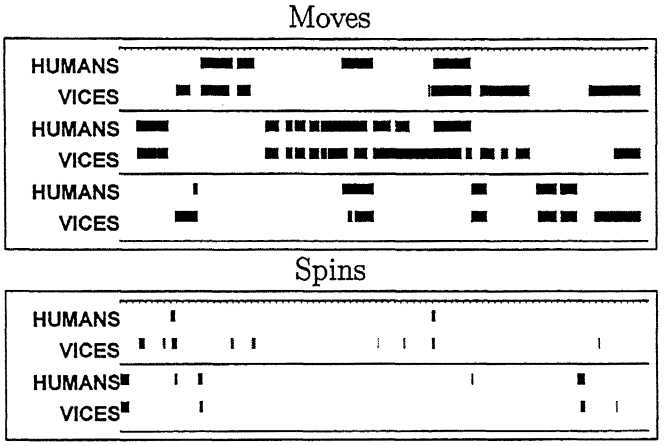


Figure 4.6: compares the descriptions between humans and VICES for monadic verbs.

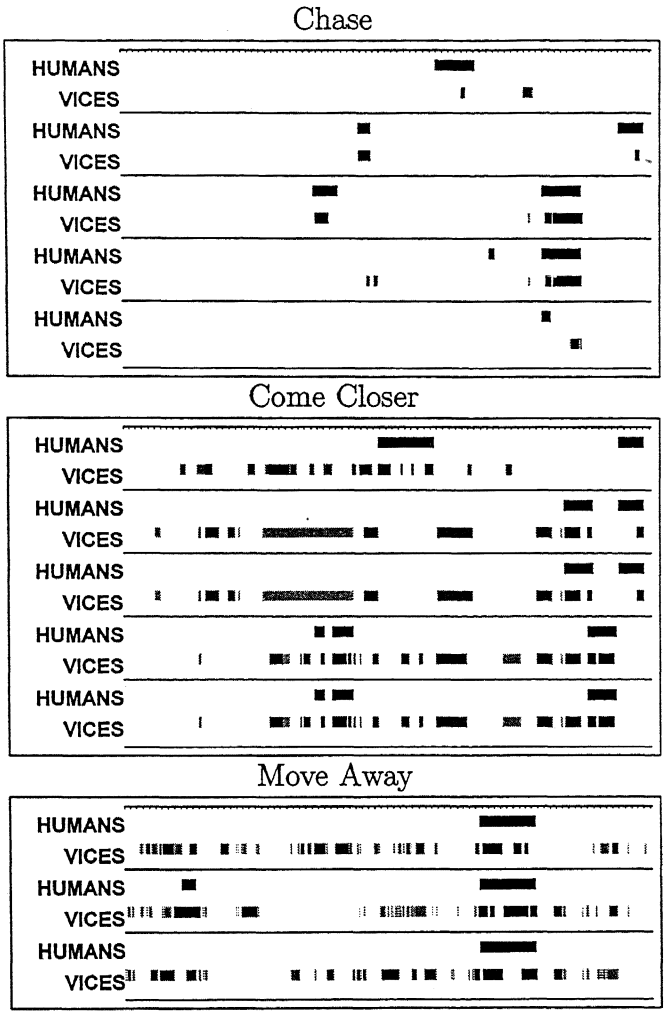


Figure 4.7: plots the performance of VICES with respect to human descriptions. The same pair of objects are chosen in these plots as are in fig 4.5

Chapter 5

Feedback Correction to reduce Focus Mismatches

We conducted an experiment to reduce focus mismatches. The *move away* verb was chosen to demonstrate that feedback correction is a tool to improve the performance of the framework with respect to focus mismatches. *Move away* is particularly affected by focus mismatches. While, human subjects do describe certain actions as *move away* actions, they also miss many instances where the objects are moving away from each other, but are overseen because of focus.

We consider the output of the system, the speech act, as an action. Comparison of this output to the expected output is then considered as a feedback to our system. In a supervised mode, these feedback are used to refine the system's accuracy, by correcting its output in case of a negative feedback. These corrections are performed through the method of bootstrapping.

We select cases where the speech act is presently incorrect and use them as explicit examples to train the SRN. Table 5.1 gives the improvement in the performance of the system after correction using feedback, for one case of the *move away* verb. Table 5.1 provides the time line comparison after feedback correction. The comparison establishes that feedback correction vastly decreased the occurrence of focus mismatches.

All measures of accuracy improve, with the true positive slightly increasing, but

the errors, specifically measured through false positives and Focus Mismatch improve drastically. The circuit was over generalizing before the feedback correction, and on correction, the over generalization cases decrease. To return to our analogy with child language acquisition, a child would learn a more coarse conceptual schema of a verb and successively refine it with further observation, specifically when supervised by an adult. The supervised learning in the feedback correction mode reflects the correction of a child's utterances by the adult.

State	True Positives	False Positives	False Negatives	Focus Mismatches	Accuracy
Before feedback correction	46.34%	7.21%	52.33%	15.95%	73.37%
After feedback correction	52.11%	1.64%	46.56%	5.88%	87.83%

Table 5.1: Improvement after feedback correction of *move away* verb

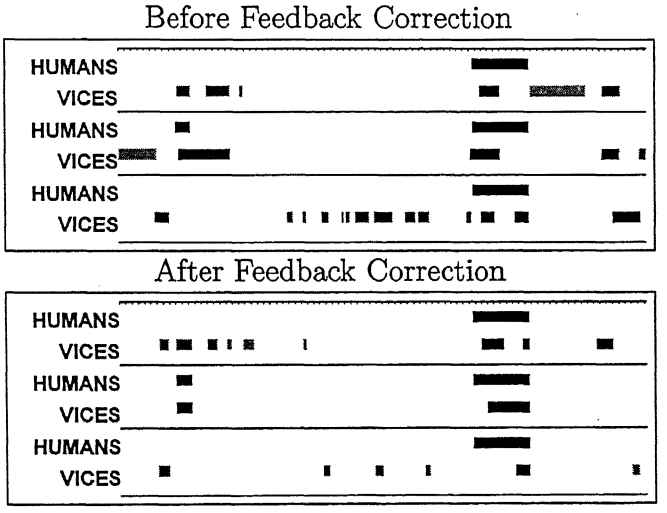


Figure 5.1: A comparison of the output of VICES before and after feedback correction for *move away* verb.

Chapter 6

Conclusion and Future work

In this work we show how a multimodal corpus can be used to learn event structures associated with action words. The data used was identical to the corpus collected by Martin and Tversky, but we are now looking to expand on the number of commentaries so that additional issues can be addressed. Another important avenue of current exploration is to generate "similar" videos, and generate mappings for these - this would also permit the learning of fine nuances between event structures of related action words.

An alternate approach, which becomes possible in the presence of additional data, would be to learn both the sensory maps and also the syntactical ramifications of the words concerned. In this work, with the limited data availability, word morphology effects such as tense, case, gender etc are assumed to be known and are ignored. The input is in text format, and a host of phonological complications such as detection of word boundaries are avoided; in the presence of speech this may also be availed. [38].

The use of SRNs, effective for small sequences, turns out to be less effective for longer event structures, e.g. in the event "playing hide-and-seek". In future, we hope to explore alternative approaches including LSTM [39] which can generalize on longer temporal histories. The set of visual features, although cognitively plausible, is currently decided manually; machine learning approaches to feature characterization need to be explored.

Another intriguing possibility that arises here is to model the act of commentary itself, where the subject, for example, has been asked to "comment on the events at a coarse level" - this speech act requires decisions to be made about actions in the scene, calling for a visual focus. In human subjects, this can be easily studied by following their gaze; in the long run, this may also lead to richer feature models, and also models of event saliency and linguistic focus.

Our framework also demonstrates a set of errors, called *Focus Mismatch*, which deviate from the commentary by human subjects, and can be attributed to focus. We also show that focus mismatch can be vastly reduced using feedback. Another avenue of future exploration would be to extend this work to 3D motion videos; having learned an action such as "chase" from a 2D video, can we extend it to recognize the same action in a 3D action video?

In closing, we would like to re-emphasize the fact that the multimodal corpus used here provides a tremendous resource for investigating language acquisition - the sterling purity of image analysis combined with multiple perspectives on the events provide an ideal environment for the task at hand. We hope to contribute to the creation of more such corpora, and also visually "similar" event corpora, which would lead to further impetus in grounded language acquisition.

Appendix A

Sample Sequences from Synthetic Video

Samples of the Synthetic Video used for training the SRNs in the second and the third framework.

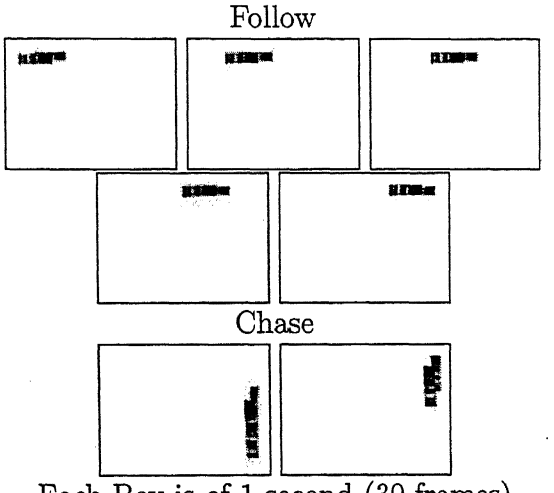


Figure A.1: Samples of sequences used to provide additional data to train VICES.

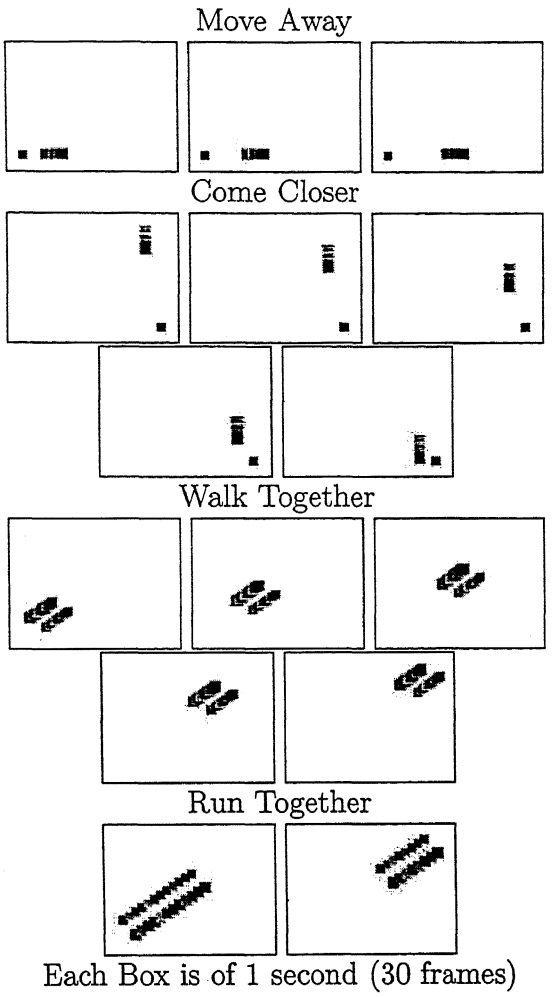


Figure A.2: Samples of sequences used to provide additional data to train VICES.

Appendix B

Resolution of Lexical Units

B.1 Original List

The original list of lexial units and the corresponding number of intervals in which they were used, for both *Chase* and *Hide and Seek* domains, are listed in table B.1, B.2. We have eliminated those predicates with frequency of just one.

Predicate	No. of Utterances	Predicate	No. of Utterances
chases	7	come together	2
displace to	4	finds	26
finds their SPOT	11	follows	2
go to	8	hide	9
looking for	3	meets at	2
move apart	4	move around	8
move away	11	move out	2
moves	32	moves to	22
playing hide and seek	7	reunite	2
rotate	3	shakes	2
spin around	3	spins	5
stop	9	trying to find	3

Table B.1: Original List of Predicates in *Hide and Seek* domain

Predicate	No. of Utterances	Predicate	No. of Utterances
breaks	2	chase	12
close	8	closes door	15
comes in	4	comes out	2
corners	3	enter	4
exit	2	follows	2
go	2	go off screen	3
go to	6	going in circle	2
hit	6	hit each other	2
leave	3	leave the box	2
move around	3	moves	7
moves to	2	open	4
opens door	7	push	4
spins	3	stop	5
stop at	2	turned	2
went inside the box	2		

Table B.2: Original List of Predicates in *Chase* Domain

B.2 Clustering on common Cognate

On repeated application of the two criterion listed in section 2.1.1 for resolving the available lexical units, a cluster of lexical units are formed as shown in fig B.1 and B.2. The lexical units, are clustered on a common cognate. The selected cognate for each of the clusters is marked in bold and is selected on the basis of the highest frequency among all lexical units in a given cluster and this cognate is the predicate considered. Each cluster gives rise to only one predicate.

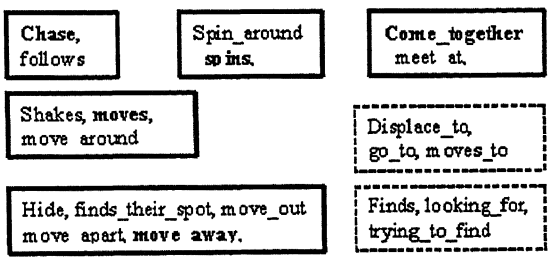


Figure B.1: Clustering of lexical units in *Hide and Seek* domain

Some of the clusters are shown in dotted, lightly boxes. These cognates have

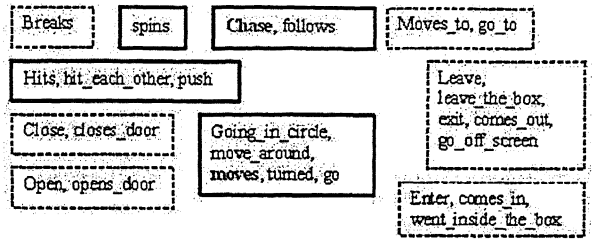


Figure B.2: Clustering of lexical units in *Chase* domain

semantics of a higher level, and need a higher level of spatial cognition, for its acquisition. Event structure of such verbs might be acquired at the stage where a better understanding of language might have been evolved. Since this study focuses on pre-linguistic acquisition, these clusters are ignored for the time being.

B.3 Final set of verbs used for the study

The remaining verbs with their semantic connotations and argument structure are given in table B.3.

Predicate	Arguments	Semantics
chases	X, Y	X is chasing Y (X behind, Y ahead)
come_together	X, Y	X and Y come together to each other
move_away	X, Y	X and Y move away from each other
hit	X, Y	X hits Y
moves	X	X moves
spins	X	X is spinning

Table B.3: Final List of Predicates

Appendix C

Feature Extraction Methodology

C.1 Static-Features Extraction

The steps involved in the analysis of each frame is as follows:

1. The image is read into a two dimension matrix, each containing the corresponding pixel value.

2. Connected Component Analysis (CCA) is performed on this. A simple, indigenously developed algorithm for CCA is used. A simple, indigenously developed algorithm is used for Connected Component Analysis (CCA). The same is outlined below:

 - In the first pass, each row in analyzed and a distinct label is assigned for each contiguous set of pixels. For fig C.1-(A), the assignment is shown in fig C.1 (B).
 - In the second pass, each column in analyzed and checked to see if two distinctly labeled regions are connected vertically. If they are, then the label assignment is merged and this forms one connected region in fig C.1 - (C).
 - In the third pass, eight connectivity check is performed to see if any two distinctly labeled regions are diagonally connected to each other. If they

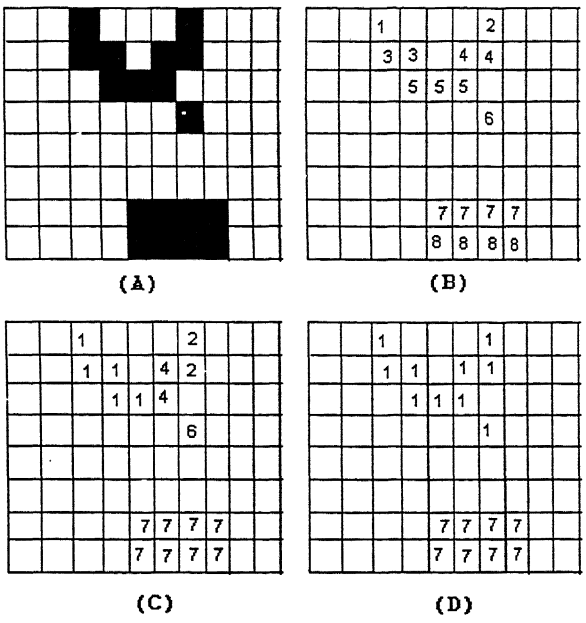


Figure C.1: Connected Component Analysis. A shows the initial set of pixels. B indicates the numbering after the first pass - horizontal assignment. C indicates the numbering after second pass - vertical assignment. D shows the final numbering after third pass - diagonal assignment.

are, then the label assignment is merged and this forms one connected region as shown in fig C.1 - (D).

This gives us each distinctly connected set of pixels. Each of them is called a blob.

3. For each blob, the bounding box (BB) is found. This is done so that all future computations may be performed on this sub region of the matrix.

4. Static features such as $x, y, area\&\theta$ are directly computed using the formulae given in the section 3. θ is computed as given in fig C.2. Three points, the bottom, the rightmost and the leftmost points are used. The angle between horizontal axis and the line joining bottom and rightmost point is called θ_1. The angle between horizontal axis and the line joining bottom and leftmost

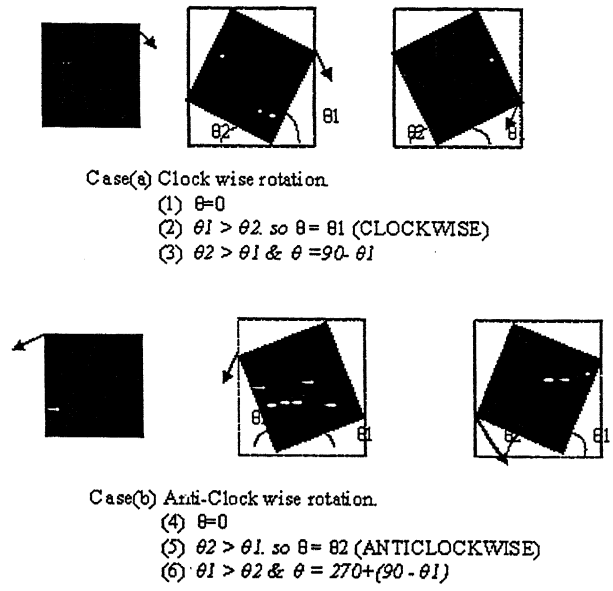


Figure C.2: Theta Computation for squares and rectangles

point is called θ_2.

Based on the slope of the line joining leftmost point and rightmost point, the direction of the rotation (clockwise or anti-clockwise) can be computed.

Based on the previous and current values of $\theta_1 \& \theta_2$, and using separate cases for each quadrant, the absolute measure of θ can be computed.

5. Tests are done to identify the type of the blob, whether it is a square, circle or a rectangle.

C.2 Temporal-Features Extraction

All objects found after analysis of each frame are called immediate objects (ImmObjs). All objects recorded so far are called Global Objects (GlobObjs). All objects found in the first frame become GlobObjs by default.

To compute the temporal features, each ImmObj is compared with all the GlobObjs. The comparison criteria is the euclidean distance and the difference in mass. If the ImmObj matches on these two criteria to any of the GlobObj, then this ImmObj is considered to be the temporal copy of this GlobObj. The temporal features of this GlobObj are then updated based on the properties of all the frames so far. All dyadic features are computed in a separate iteration after all the monadic features have been computed.

C.3 Dynamic Tracking system to handle Occlusion, Contact and New Entries

Since the videos used in this study involves contact or touching between objects, and occlusion between objects, our system has a Dynamic Tracking system to detect these special cases and to maintain consistency in their presence.

The Dynamic Tracking system involves maintaining a track box of each object. A track box is area in which the probability of the existence of this object is maximum. The track box is illustrated in fig C.3. The dotted lines indicate the track box. The maximum distance that can be covered by this object is computed as $d = vt + \frac{at^2}{2}$. Here v and a are the latest velocity and acceleration as maintained in the GlobObj properties. This distance in all four directions give us the bounds of the track box relative to the present position.

Whenever the track Boxes of any two objects overlap (as in fig C.3), we conclude that the probability of a touch or occlusion is high. When this happens, a finer level analysis is performed. This involves finding out all those GlobObjs for which a corresponding ImmObj was not found. All such GlobObjs are collected as classified as Missing Objects (MissObjs). Similarly, there will be a set of ImmObjs which has not been classified as any of the existing GlobObjs. These are called Extra Objects (ExtObjs). Each combination of two MissObjs are taken at a time and compared to the ExtObjs on the basis of combined mass and euclidean distance. If a match is found, then it is assumed that the two objects are touching or occluding. Once this is found, all the features are computed accordingly. In case of fig C.3, finer analysis

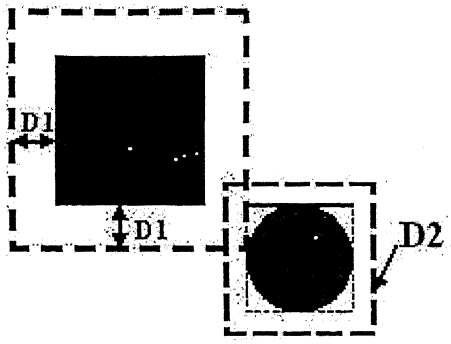


Figure C.3: Track Box for Dynamic Tracking

will lead to the conclusion that they do not touch or overlap.

All ExtObjs not classified as any combination of touching or occluding objects are considered to be objects newly entered into the video at this instance of time. New GlobObjs are created to handle them.

Bibliography

[1] ASUDEH, A. Neural constructivism and language acquisition. In *Proceedings of Electronic Conference: The 40-th Anniversary of Generativism* (December 1997).

[2] BALLARD, D. H., AND YU, C. A multimodal learning interface for word acquisition. In *IEEE international Conference on Acoustics, Speech and Signal processing (ICASSP'03)* (2003).

[3] BRAND, M., OLIVER, N., AND PENTLAND, A. Coupled hidden markov models for complex action recognition. *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)* (June 1997).

[4] BUDANITSKY, A. Lexical emantic relatedness and its application in natural language processing. Tech. rep., Computer Systems Research Group, University of Toronto, 1999.

[5] ELMAN, J. L. Finding structure in time. *Cognitive Science 14* (1990), 179–211.

[6] ELMAN, J. L. Generalization, simple recurrent networks, and the emergence of structure. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, M. Gernsbacher and S. Derry, Eds. Lawrence Erlbaum Associates, Mahwah NJ, 1998.

[7] ELMAN, J. L. Origins of language: A conspiracy theory. In *The emergence of language*, B. MacWhinney, Ed. Lawrence Erlbaum Associates, Hillsdale, NJ, 1999.

[8] FERN, A. P., GIVAN, R. L., AND SISKIND, J. Learning temporal, relational, force-dynamic event definitions from video. In *Proceedings of AAAI* (2002), pp. 159–166.

[9] FIROIU, L., AND COHEN, P. R. Segmenting time series with a hybrid neural networks - hidden markov model. *Eighteenth national conference on Artificial intelligence* (2002), 247–252.

[10] GIESE, M. A. Neural field model for the recognition of biological motion patterns. In *Proceedings of NC 2000* (May 2000).

[11] GIESE, M. A. Neural model for the recognition of biological motion. In *Dynamische Perzeption*. Infix Verlag, 2000, pp. 105–110.

[12] GIESE, M. A., AND POGGIO, T. Neural mechanisms for the recognition of biological movements. *Nature, Neuroscience 4* (March 2003), 179–192.

[13] HEIDER, F., AND SIMMEL, M. An experimental study of apparent behavior. *Americal Journal of Psychology 57* (1944), 243–259.

[14] HOWELL, A., AND BUXTON, H. Recognising simple behaviours using time-delay rbf networks. *Neural Processing Letters 5*, 2 (April 1997), 97–105.

[15] KADOUS, M. W. *Temporal Classification: Extending the Classification Pradigm to Multivariate Time Series*. PhD thesis, School of CSE, University of New South Wales, October 2002.

[16] KAMIMURA, R. Application of the recurrent neural network to the problem of language acquisition. In *Proceedings of the conference on Analysis of neural network applications* (1991), ACM Press, pp. 14–28.

[17] KRIPKE, S. *Wittgenstein on Rules and Private Language*. Oxford University Press, 1982.

[18] LI, X., AND PORIKLI, F. Traffic event detection: A survey, 2003.

[19] MANDLER, J. M. How to build a baby ii. conceptual primitives. *Psychological Review 99* (October 1992), 587–604.

[20] MARTIN, B., AND TVERSKY, B. Segmenting ambiguous events. In *Proceedings of the 25th annual meeting of the Cognitive Science Society* (2003). Crucial for our Data-Collection chapter.

[21] MCKENNA, S., AND GONG, S. Gesture recognition for visually mediated interaction using probabilistic event trajectories. In *Proceedings of British Machine Vision Conference 1998* (1998).

[22] MICHNICK-GOLINKOFF, R., AND HIRSH-PASEK, K. Reinterpreting children's sentence comprehension: Toward a new framework. In *Handbook of Child Development* (1995), P. Fletcher and B. MacWhinney, Eds., Blackwell, pp. 430–461.

[23] MOORE, D. J., AND ESSA, I. A. Recognizing multitasked activities from video using stochastic context-free grammar. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence* (July-August 2002), pp. 770–776.

[24] MOZER, M. C. Neural network architectures for temporal pattern processing. *Time series prediction: Forecasting the future and understanding the past 16* (1993), 243–264.

[25] OATES, T. Peruse: An unsupervised algorithm for finding recurring patterns in time series. In *2002 IEEE International Conference on Data Mining (ICDM'02)* (2002), p. 330.

[26] OATES, T., COHEN, P. R., ATKIN, M. S., AND BEAL, C. R. Building a baby. In *Proceedings of the 18th annual conference of the Cognitive Science Society* (1996), pp. 518–522.

[27] OLIVER, N. M., ROSARIO, B., AND PENTLAND, A. P. A bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 8 (2000), 831–843.

[28] PARK, S., AND AGGARWAL, J. Recognition of two-person interactions using a hierarchical bayesian network. In *ACM SIGMM 2003 Workshop on Video Surveillance* (2003).

[29] PINHANEZ, C. *Representation and recognition of action in interactive spaces.* PhD thesis, MIT Media Lab, 1999.

[30] PLUNKETT, K. Connectionist approach to language acquisition. In *Handbook of Child Development* (1995), P. Fletcher and B. MacWhinney, Eds., Blackwell, pp. 36–72.

[31] QUINE, W. V. O. *Word and Object.* New York: John Wiley and Sons, Cambridge: MIT, 1960.

[32] RAO, R. P. Bayesian computation in recurrent neural circuits. *Neural Computation 16* (2004), 1–38.

[33] REGIER, T. *The Human Semantic Potential.* MIT Press, Cambridge, MA, 1996.

[34] ROY, D. K. Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language*, 16 (July-October 2002), 353–385.

[35] ROY, D. K., AND PENTLAND, A. P. Learning words from sights and sounds: a computational model. *Cognitive Science 26* (January/February 2002), 113–146.

[36] RUEPING, S., AND MORIK, K. Support vector machines and learning about time. Tech. rep., CS, Dortmund, 2003.

[37] RUPLING, S. Svm kernels for time series analysis. In *LLWA 01 - Tagungsband der GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivity* (2001), pp. 43–50.

[38] SAFFRAN, J. R., SENGHAS, A., AND TRUESWELL, J. C. The acquisition of language by children. *PNAS 98*, 23 (November 2001), 12874–12875.

[39] SCHMIDHUBER, J., GERS, F., AND ECK, D. Learning nonregular languages: A comparison of simple recurrent networks and lstm. *Neural Computation 14*, 9 (2002), 2039–2041.

[40] SEBE, N., COHEN, I., GARG, A., LEW, M., AND HUANG, T. Emotion recognition using a cauchy naive bayes classifier. In *Proceedings of International Conference on Pattern Recognition (ICPR'02)* (August 2002), pp. 17–20.

[41] SISKIND, J. Axiomatic support for event perception. In *Proceedings of the AAAI Workshop on Integration of Natural Language and Vision Processing* (August 1994), pp. 153–160.

[42] SISKIND, J. Visual event perception. In *Symbolic Visual Learning*, K.Ikeuchi and M. Veloso, Eds. Oxford University Press, NY, 1996.

[43] SISKIND, J. M. Grounding language in perception. *AI Review 8*, 5-6 (1995), 371–391.

[44] SNOEK, C., AND WORRING, M. Time interval based modelling and classification of events in soccer video. In *Proceedings of the 9th Annual Conference of the Advanced School for Computing and Imaging (ASCI)* (June 2003).

[45] STEELS, L. Evolving grounded communication for robots. *Trends in Cognitive Sciences 7*, 7 (July 2003), 308–312.

[46] TARABORELLI, D. What is a Feature? A Fast and Frugal Approach to the Study of Visual Properties. In *Proceedings of the Eighth International Colloquium on Cognitive Science* (2003).

[47] TOWSEY, M., DIEDERICH, J., SCHELLHAMMER, I., CHALUP, S., AND BERGMAN, C. Natural language learning by recurrent neural networks: a comparison with probabilistic approaches. In *New Methods in Language Processing and Computational Natural Language Learning*, D. Powers, Ed. ACL, 1998, pp. 3–10.

[48] XINGHUA, S., GUOYING, J., MEI, H., AND GUANGYOU, X. Bayesian network based soccer video event detection and retrieval. In *Proceedings of SPIE - The International Society for Optical Engineering* (2003), pp. 839–842.

[49] ZACKS, J. M., TVERSKY, B., AND IYER, G. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology 130*, 1 (2001), 29–58.